SVD-GAN for Real-Time Unsupervised Video Anomaly Detection

BMVC 2021 Submission # 1295

Abstract

Real-time unsupervised anomaly detection from videos is challenging due to the uncertainty in occurrence and definition of abnormal events. To overcome this ambiguity, an unsupervised adversarial learning model is proposed to detect such unusual events. The proposed end-to-end system is based on a Generative Adversarial Network (GAN) architecture with spatiotemporal feature learning and a new Singular Value Decomposition (SVD) loss function for robust reconstruction and video anomaly detection. The loss employs efficient low-rank approximations of the matrices involved to drive the convergence of the model. During training, the model strives to learn the relevant normal data distribution. Anomalies are then detected as frames whose reconstruction error, based on such distribution, shows a significant deviation. The model is efficient and lightweight due to our adoption of depth-wise separable convolution. The complete system is validated upon several benchmark datasets and proven to be robust for complex video anomaly detection, in terms of both AUC and Equal Error Rate (EER).

1 Introduction

Anomalies can be defined as unusual events which deviate from 'normal' behaviour. Anomaly detection is relevant to various applications, including intrusion detection in time series, surveillance, action detection and healthcare monitoring. Manually detecting anomalies from surveillance videos is a painstaking job, and a demanding one in terms of human resources. With the increase of surveillance cameras the need for an automated system for anomaly detection from videos has thus gained much recognition. These systems play a crucial role in security control, crime detection, accidents and traffic monitoring, where the data available from various sources is simply too much for manual analysis.

The occurrence of an anomalous event is unexpected and rare in practice, making it difficult to categorise such diverse events. For instance, people running on the beach is considered normal behaviour, whereas running in shopping mall is considered anomalous. Therefore, automated system will not have any prior knowledge about the nature of past or future anomalies. Representational learning methods, such as sparse-coding, achieve good performance in anomaly detection [20, 46]. Anomaly detection in videos, in particular, can be posed in either the supervised or the unsupervised learning setting. In supervised anomaly detection, the system learns from example videos labelled as either anomalous or non-anomalous (i.e. normal) [17, 37, 45]. In the unsupervised approach, instead, the system

assumes any rare or abnormal occurrence, deviating from the learned normal sample parameters, to be anomalous [33]. The model can thus be trained to detect anomalies using huge volumes of unlabeled, 'normal' data. In surveillance, variations due to e.g. changes in scale and viewpoint introduce additional ambiguity.

A number of studies have focused on the use of Convolutional Neural Networks (CNNs) for (supervised) anomaly detection in industrial products [43], for instance inspecting cement surfaces [6] and cracks [15]. In the supervised setting, however, the issue arises of an often uneven balance between normal and abnormal data. The use of data augmentation has been proposed to mitigate this challenge – nevertheless, this setting still has serious limitation in addressing real life problems [9, 47]. In opposition, Ravanbakhsh et al. have more recently proposed to employ adversarial learning to localise anomalous activities [30] in an *unsupervised* setting. Generative Adversarial Networks (GANs), in particular, as they have the capability to model high-dimensional image data, have become the state of the art in anomaly detection in recent times. Many GAN-centric architectures, such as AnoGAN [35] and GANomaly [2], have been put forward to this purpose. The objective of GAN, however, which encourages the generated samples to look like real data, is not directly aligned to the goal of performing anomaly detection. Consequently, in much recent work in anomaly detection adversarial training has been modified to improve both training and inference for this specific goal [14, 39]. Overall, both CNN-based and GAN-based architectures are inefficient to run on edge devices such as robots, smart surveillance cameras, autonomous driving cars or microcomputers. In addition, most of the networks used in unsupervised GAN architectures are shallow and are designed to learn only spatial features, ignoring the crucial temporal component of videos. As a result, such networks are tapered to low-dimensional data and prone to overfitting because of the large number of parameters.

Contributions. In this paper, inspired by the GANomaly [2] architecture but significantly departing from it, we propose a light-weight, efficient anomaly detection architecture with a reduced number of parameters, aimed at addressing the convergence and overfitting problems with GAN training and at achieving real-time performance. Our SVD-GAN has an encoder-decoder architecture as the backbone, and is capable of learning in an unsupervised way both spatial and temporal features from real-time videos, thanks to a stacked convolutional Long-Short Term Memory (LSTM) network structure. Our main contributions are:

- *An original Singular Value Decomposition (SVD) loss function* which backpropagates only a small number of 'dominant' patterns from the input and generated video frames as loss value, thus improving the reconstruction ability of the generator.

- *An original generator structure* making use of *depth-wise convolution* layers of our own design, which leads to increased efficiency due to a reduction in the number of model parameters of 15.9% without compromising on feature extraction, with the net effect of a model that is both lightweight and more stable.

Whereas using the proposed SVD loss in GANs drastically improves their performance (as it minimises the KL divergence between real and generated data distributions), the new loss can be widely employed in other deep learning models for better representation learning, whenever few training samples are available (e.g., in few shot learning). To avoid the vanishing gradients issue, network parameters are updated via distinct sub gradients [12] using residual connections, in an original architecture for spatiotemporal feature extraction from videos. Our system can detect complex anomalous events occurring for a very short time, and outperforms the prior art on the large-scale CHUK Avenue and ShanghaiTech datasets.

2 Background

Deep learning for anomaly detection. Deep learning-based methods have achieved success in detecting abnormal events from videos, outperforming the former state of the art in challenging environments [7, 19, 21, 33, 40]. Deep neural networks with hierarchical feature representation learning are simply more powerful than the hand crafted feature extraction techniques used in traditional architectures. Deep generative models, in particular, have recently come to the fore with their ability to encode complex transformations. Liu et al. proposed the use of GANs to detect anomalies by minimising the difference between predicted future video frames and ground truth frames [19]. During training the normal data distribution is estimated using the training video frames $\mathbf{X}_{train} = \{X_i\}$ by learning a parametric representation $f_\theta : \mathbf{X}_{train} \rightarrow \mathbb{R}^m$ which minimises the model reconstruction loss. At test time, an anomaly score $A(X_j)$ is computed for each test frame $X_j \in \mathbf{X}_{test}$ as the deviation from the learnt optimal representation f_{θ^*}: $A(X_j) = \left\| f_{\theta^*}(X_j) - X_j \right\|^2$. Finally, anomalies are detected by applying a threshold T, $A(X_j) > T$, to the anomaly score.

Generative Adversarial Networks consist of two different networks, a generator and a discriminator, both trained with unlabeled data [10, 29, 32]. The generator G aims to capture the data distribution and generate realistic video frames, by building a data distribution for the input data X via a mapping from a prior latent space noise distribution z. The objective of the discriminator D, instead, is to find the probability of the sample being outputted by the generator. Generator and discriminator compete against each other, by playing a zero-sum min-max game: $\min_G \max_D V(D,G) = E_{X \sim p_{data}(X)} \log D(X) + E_{Z \sim p_z(Z)} \log(1 - D(G(Z)))$.

In recent years, various anomaly detection GAN architectures have been proposed. One such architecture is the extended conditional GAN proposed by Mizra et al. [25]. This model conditions either the generator G or the discriminator D using some additional information Y. The condition Y can be formulated from multimodal input data or class labels available as auxiliary information. Vu et al. [41] have proposed a robust anomaly detection system for videos which uses conditional GANs to detect video anomalies accurately at various levels of representation using a layer-wise approach. The extraction of optical flow data poses challenges in uncertain environments characterised by untextured regions, illumination changes, occlusions and fast motions. The Bidirectional GAN (BiGAN) architecture proposed by Donahue et al. [8] consists of an encoder $E\ (X, E(X))$ which learns the inverse mapping of the generator, $E = G - 1$. During training, the model learns how to map the latent space to the image data and vice-versa, reducing the statistical complexity and producing better results on the MNIST benchmark dataset [8].

In 2019, a new mapping scheme called 'fast-AnoGAN' was proposed by Schlegl et al. [34], which is capable of fast detection of anomalies at image level and pixel-level localisation. In 2018, Zenati et al. proposed an Efficient GAN-Based Anomaly Detection (EGBAD) system which uses the BiGAN architecture [44]. Ackay et al., instead, hypothesized the learning of image space and latent space vectors jointly. The architecture employs an adversarial autoencoder with an encoder-decoder-encoder pipeline for capturing the data distribution of image and latent space vectors. This architecture, however, is limited in the way it handles spatial-temporal learning, and produces unstable reconstructions for real-time videos [2]. Vu et al. proposed a robust anomaly detection system for videos which uses representational learning from both intensity and motion information via conditional GANs [41]. Nguyen and Meunier [27] proposed a deep CNN that addresses anomaly detection by learning a correspondence between common object appearances (e.g. pedestrian, background,

tree, etc.) and their associated motions. A combination of future frame prediction and reconstruction for anomaly detection was proposed by Tang et al. using two U-Net blocks in the generator [1], where the first block tries to predict frames while the other reconstructs the frames generated by the former block.

Convolutional LSTMs. Jefferson et al. proposed a Conv-LSTM network in an encoder-decoder model for the prediction of future frames and anomaly detection by reconstruction [24]. The same architecture was proved to be promising for video anomaly detection by [24] and [22]. Input video frames are there passed to the convolutional LSTM for feature extraction and then reconstructed using deconvolution. Luo et al. proposed a Temporally-coherent Sparse Coding (TSC) approach in which similar neighbouring frames are mapped via stacked Recurrent Neural Networks (RNNs) to a reconstruction coefficient [21]. LSTM autoencoders are suited to extracting spatial-temporal information. Shi et al. [36] and Patraucean et al. [28] both used stacked convolutional LSTMs in an autoencoder architecture for feature extraction in video sequence data. Conv-LSTMs can capture spatial representations from the video frames and improve the ability to predict future frames, while showing a strong ability to characterise spatio-temporal information.

In Conv-LSTMs the amount of information from the previous time step received by the hidden state is partly determined by the size of convolutional filter in the hidden-to-hidden connection. To capture faster motions large transitional kernels are used, while for slower motions small kernels can do [41]. GAN performance, however, degrades drastically as the number of parameters increases, leading GANs to often fail on high dimensional data. To overcome this, in our model *depth-wise separable convolutions* are used within the LSTM, reducing the size of the model and the chance of overfitting during GAN training. Depth-wise separable convolution is followed by pointwise convolution to deal with the spatial and depth dimensions of the video frames, thus splitting a 3×3 kernel into 3×1 and 1×3 kernels. Input frame has three separate filters for R, G and B channels.

3 Architecture and methodology

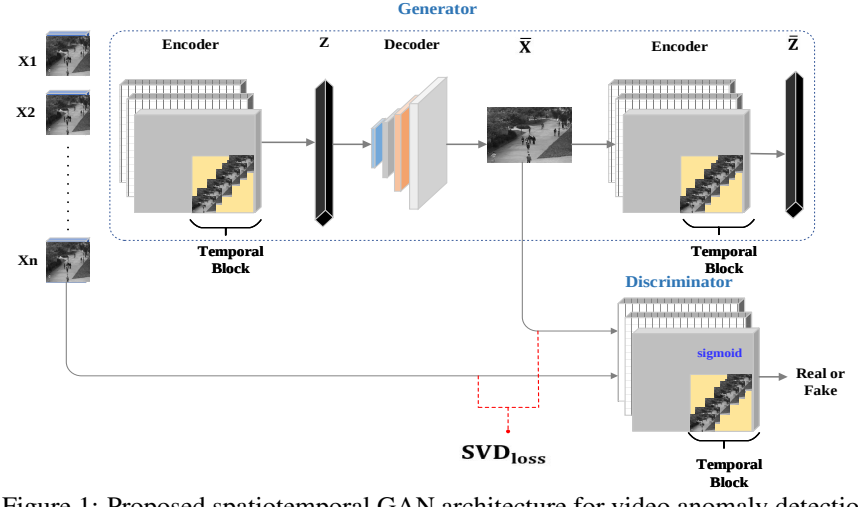

Figure 1: Proposed spatiotemporal GAN architecture for video anomaly detection.

Given a number of unlabeled sample video frames $\mathbf{X}_{train} = (X_1, \ldots, X_n)$ at training time, our unsupervised anomaly detection approach is designed to learn suitable spatiotemporal

features from $\mathbf{X}_{train}$ that, at test time, it employs to detect whether a new video frame X is anomalous via an anomaly score $A(X)$. The models learn from $\mathbf{X}_{train}$ the probability distribution of the 'normal' frames while minimising the anomaly score of the training samples.

3.1 Architecture

Our proposed architecture (Figure 1) is based on the generative adversarial network principle and uses an encoder-decoder-encoder [2] pipeline as the Generator, which learns feature representations directly from the input samples, and an encoder-based Discriminator which aims to discern real from fake images.

At each time step, a batch of video frames of fixed duration is passed as input to both the generator G and the discriminator D. Each input image X is passed to the encoder E in the generator G, which maps it to a latent space $Z = E(X)$. The decoder then transposes these latent space vectors back to the image data space, implementing a mapping $\bar{X} = G(X)$.

The first contribution of this paper is an original Generator architecture, shown in Figure 2, where it can be seen how the generator learns the temporal dependencies within a video sequence using temporal blocks of frames. The architecture is fully described in Sec. 3.4.

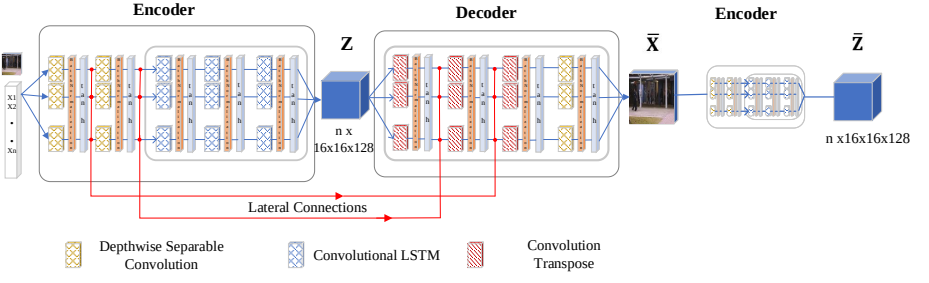

Figure 2: Pipeline of our spatiotemporal GAN generator (full description in Sec. 3.4).

At training time, the network learns the joint posterior distribution of the data $G(Z,X)$, and each input sample X is encoded using its latent representation. Training is performed using the inverse mapping from image data to latent space proposed by Lipton et al [13]. The generated image $\bar{X} = G(X)$ and the input image X are then passed to the discriminator, which discriminates real from generated frames using a sigmoid activation function and backpropagates the loss to the generator itself.

3.2 Singular Value Decomposition (SVD) loss

GANs are known to have convergence issues with high-dimensional input data (e.g., video frames). Vanishing gradient problems tend to occur in the generator during training.

Minimising the KL divergence. GAN training, however, can be improved by minimising the Kullback-Leibler (KL) divergence between the 'real' data distribution $\mathbb{P}_{data}(X)$ and the 'generated' distribution $\mathbb{P}_{gen}(X) \doteq \mathbb{P}(G(X))$, i.e., the probability distribution of the generated images [3]:

$$KL(\mathbb{P}_{data} \parallel \mathbb{P}_{gen}) = \int_X \mathbb{P}_{data}(X) \log \frac{\mathbb{P}_{data}(X)}{\mathbb{P}_{gen}(X)} dX = E_{X \sim \mathbb{P}_{data}(X)}[\log \mathbb{P}(X) - \log \mathbb{P}(G(X))].$$

Making the generated distribution as close as possible to the real one is indeed the main objective of GAN. Various measures of dissimilarity as possible, including the total variation and the Wasserstein distances [3].

In this paper, rather than by posing as an objective the minimisation of the KL divergence of the distributions of real and generated data, we aim to achieve this through an original Singular Value Decomposition (SVD) loss function designed to backpropagate the difference between low-rank approximations of input X and generated images $G(X)$ defined by their principal components.

Properties of SVD decomposition. SVD is a matrix decomposition technique that, when applied to images, is able to encode the maximum fraction of signal energy using only a few coefficients [4, 26, 31]. An input image matrix X (e.g., a video frame) with m rows and n columns factorises as:

$$X = USV^T$$

where U is an orthogonal $m \times m$ matrix, V is orthogonal and $n \times n$, and S is a matrix of the same size as X, with on the diagonal the singular values σ_i of X (see Fig. 3). The columns u_i of U are the eigenvectors of XX^T, while the columns v_i of V are the eigenvectors of X^TX. A fundamental property of SVD is that the matrix $\widehat{X}_r = \sum_{i=1}^{r} \sigma_i u_i v_i^T$ (with the eigenvalues in order of magnitude) is the optimal solution to the problem $\min \|X - X'\|_F$ (where F denotes the Frobenius norm) subject to $\text{rank}(X') \leq r$, i.e, is the optimal rank-r approximation of X.

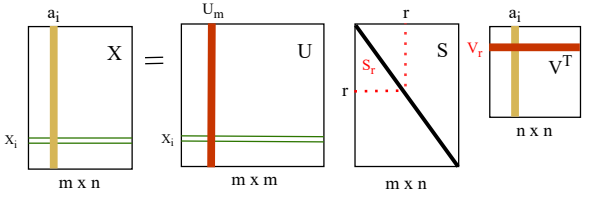

Figure 3: Graphical representation of the SVD decomposition of a matrix (frame) X.

SVD loss. Our proposal is to minimise the empirical expectation of the L2 norm of the difference between the low-rank SVD approximations of the input image X and of the generated image $G(X)$, which we term *SVD loss*:

$$SVD_{loss} = \mathbb{E}_{X \sim \mathbb{P}(X)} |\widehat{X}_r - \widehat{G(X)}_r|_2, \tag{1}$$

where $\widehat{X}_r$ and $\widehat{G(X)}_r$ are the rank-r approximations of X and $G(X)$, respectively.

Our conjecture, which we empirically validate in this paper, is that minimising the SVD loss is indeed correlated with minimising the KL divergence between real and generated data, and should thus have positive effects on the convergence of our model.

3.3 Overall loss

In our model, more stable GAN training is achieved by using as overall loss:

$$SVD_{loss} + ADV_{loss} + CONT_{loss} + ENCOD_{loss}. \tag{2}$$

The *adversarial loss* ADV_{loss} is the Jensen-Shannon divergence of the output of the discriminator D for the input image X and the corresponding generated image $G(X)$ [32], namely: $ADV_{loss} = \mathbb{E}_{X \sim \mathbb{P}(X)} |D(X) - D(G(X))|$. A *contextual loss* based on the L1 norm is used for measuring the distance between the input image X and reconstructed image $G(X)$, penalizes G as [13]: $CONT_{loss} = \mathbb{E}_{X \sim \mathbb{P}(X)} |X - G(X)|_1$. Our SVD loss can thus be seen as a an efficient form of contextual loss, based on the L2 norm. Compared to the L1 norm, L2 better penalises

outliers, as deviations are magnified by taking the square. Finally, the *encoder loss* minimises the distance between the bottleneck features of the input $Z = E(X)$ and the encoded features of the generated image $\bar{Z} = E(G(X))$, namely: $ENCOD_{loss} = \mathbb{E}_{X \sim \mathbb{P}(X)} |Z(X) - \bar{Z}(X)|_2$.

3.4 Generator structure and training protocol

We trained our spatiotemporal GAN model on an 8-GPU machine with Quadro RTX 6000 cards having 24 GB VRAM each. Input frames were resized to 128×128 pixels and passed to the SVD-GAN architecture. The proposed architecture uses *tanh* activations in the generator and LeakyRelu in the discriminator. Batch normalisation with *tanh* at the end of each layer helps scaling and adjusting the input features between -1 and 1. The generator uses an Adam optimiser with first order derivative. The discriminator uses RMSProp with a 0.00005 learning rate for weight optimisation.

In our Generator, each batch of n rescaled input frames goes through two layers of depthwise convolution and 4 layers of convolutional LSTM for spatio-temporal feature learning, as shown in Figure 2. Frames are convolved with a kernel of size 5×5 and stride 2 to produce a feature map of size $n \times 64 \times 64 \times 128$. Subsequently, a small kernel of size 3×3 is applied to the respective feature maps to capture spatiotemporal features using a block of convolutional LSTMs. Input frames are encoded to a latent space Z of size $n \times 16 \times 16 \times 128$ to be then passed to the decoder for reconstruction. The decoder uses convolutional 2D transpose layers and batch normalisation to decode the bottleneck features back to the image space ($\bar{X}$). The reconstructed data is remapped to the latent space ($\bar{Z}$) for a consistent comparison between Z and $\bar{Z}$. Finally, the generated image $\bar{X}$ and the input image X are given as inputs to the discriminator which has the same encoder architecture as the generator, with an additional sigmoid activation for discrimination. Losses are back propagated to the generator for an accurate reconstruction of the input image X.

3.5 Anomaly detection

During testing, anomalies are detected by thresholding an anomaly score $A(X)$, the square $L2$ distance between input and reconstructed images, rescaled to the interval $[0,1]$:

$$A(X) = \frac{1}{p} \sum_{i=1}^{p} (\bar{X}_j(i) - X_j(i))^2, \qquad \hat{A}(X) = \frac{A(X) - A_{\min}}{A_{\max}},$$

where X_j is the input test video frame, $\bar{X}_j$ is the reconstructed video frame, p the number of pixels in a frame, and $A_{\min}$ / $A_{\max}$ are its minimum / maximum over the test sequence.

Note that we do not commit to any specific threshold, but assess the performance of a model over the whole range of thresholds by measuring the Area Under the ROC Curve (AUC), after plotting the model's Receiver Operating Characteristics (ROC) curve (see Section 4.1 - Metrics).

4 Experiments

4.1 Datasets and evaluation metrics

We validate our approach over several benchmark datasets portraying complex anomalous events in various scenarios involving multiple scenes captured from different angles. All

datasets comprise 'normal' video frames for training and a combination of anomalous and non-anomalous frames for testing. Their features are summarised in Table 1.

The *CHUK Avenue dataset* contains 16 normal videos for training and 21 videos for testing, for a total of 30,652 frames [20]. Test videos include anomalies like the throwing of objects, walking in the wrong direction, running, and loitering. The *UCSD anomaly detection dataset* contains surveillance videos of pedestrian walkways [23]. Anomalies include presence of skaters, bikers, small carts and people walking sideways in walkways. The dataset is divided into two parts: Ped1 and Ped2. Ped1 contains 34 normal video samples for training with some perspective distortion and 36 videos samples for testing. Ped2 portrays pedestrians walking parallely to the camera plane, with 16 videos for training and 12 for testing. The complex *ShanghaiTech Campus* data set is specifically used to validate the robustness of the model [19]. The corpus contains video sources from 13 different scenes, under various lighting conditions and camera angles. It has 330 video samples for training and 107 for testing, with a total of 130 abnormal events. Anomalies include complex unusual human behaviors, presence of unusual objects and movements in the wrong direction.

Benchmark Datasets	**Anomalous Events**	**Sour -ces**	**Normal Frames**	**Anomalous Frames**	**Training Frames**	**Testing Frames**
UCSD Ped1	40	1	9995	4005	6800	7200
UCSD Ped2	12	1	2924	1636	2550	2010
CHUK Avenue	47	1	26832	3820	15328	15324
ShanghaiTech	130	13	300308	17090	274515	42883

Table 1: Characteristics of the datasets used in this work.

Metrics. The model's frame-level performance is analysed using the Area Under the ROC Curve (AUC), after plotting the Receiver Operating Characteristics (ROC) curve. The latter plots the True Positive Rate (TPR) vs the False Positive Rate (FPR) as a function of the detection threshold in the range [0,1], thus summarising the trade-off between TPR and FPR for a predictive model using different probability thresholds.

AUC measures the two-dimensional area under the entire ROC curve from (0,0) to (1,1), providing an aggregate measure of performance across all possible detection thresholds which amounts to a sort of probability distribution over the range of thresholds. The AUC thus represents the degree of separability the model can enforce between anomalous and non-anomalous frames. The higher the AUC value, the better the performance.
The EER (Equal Error Rate), the point on the ROC curve where false positive and false negative rates coincide, is also reported (see Table 2).

4.2 Comparison with state of the art

The performance of our SVD-GAN on all datasets is compared with that state-of-the-art *unsupervised* anomaly detection systems in Table 2. The table also compares the performance of our proposed architecture with and without SVD loss. The architecture with SVD loss achieves state-of-the-art results, by a very large margin, on both the Avenue and the ShanghaiTech datasets, with an AUC of 89.82 % and EER of 21.55% on CHUK Avenue and an AUC of 78.42% and EER of 25.16% on ShanghaiTech. The latter, in particular, compared to all other datasets is a truly large scale benchmark with 130 very complex anomalous events.

Detailed information on the number of frames and anomalous events for each dataset is provided in the supplementary material. On UCSD our model's AUC is low compared

	UCSD Ped1		UCSD Ped2		Avenue		ShanghaiTech	
Unsupervised Methods	**AUC**	**EER**	**AUC**	**EER**	**AUC**	**EER**	**AUC**	**EER**
MPPC+SFA [2010] [23]	66.8	–	61.3	–	–	–	–	–
Conv-AE [2016] [10]	81.1	27.9	90.0	21.7	70.2	25.1	–	–
ConvLSTM-AE [2017] [22]	75.5	–	88.1	–	77.0	–	–	–
sRNN [2017] [21]	–	–	92.21	–	81.71	–	68.00	–
TSC [2017] [21]	–	–	91.03	–	80.56	–	67.94	–
STAN [2018] [16]	82.1	–	96.5	–	87.2	–	–	–
Flownet+UNet [2017] [19]	83.1	–	95.4	–	85.1	–	72.8	–
MLAD [2019] [31]	82.34	**23.50**	**99.21**	**2.49**	52.82	38.82	–	–
ITAE+NFs [2020] [6]	–	–	97.3	–	85.8	–	74.7	–
ROADMAP [2021] [32]	**83.4**	–	96.3	–	88.3	–	76.6	–
Ours without SVD loss	70.53	31.20	74.6	25.32	82.93	24.56	71.75	30.24
Ours with SVD loss	73.26	28.75	76.98	23.46	**89.82**	**21.55**	**78.42**	**25.16**

Table 2: Performance comparison in terms of AUC and EER among state-of-art unsupervised anomaly detection architectures, including ours.

to the state of the art. This is likely due to the relatively short duration (ca 200 frames) of the available training sequences. In this situation the model tends to reconstruct well the anomalous frames too and falls short when detecting certain anomalies. E.g., the AUC is comparatively low for videos portraying skaters or cyclists in pedestrian pathways – these anomalies look closer to normal from the angle (top view) from which the video is captured. This behaviour is to be expected, for LSTM-based models need sufficient sequential data to be excited and perform well. A graph plotting the regularity score vs the time stamp for a test video from Shanghai Tech and Avenue datasets are shown in the supplementary material.

The variance of the processing time of frames from the mean is 4ms, and the worst-case performance (per frame) is 27ms. The overall Floating Point Operations (FLOPs) of our SVD-GAN is also relatively low, at 2.214042×10^6 FLOPs.

4.3 Overfitting and accurate reconstruction

The proposed tackles two key issues preventing a GAN-based architecture from reliably reconstructing video frames. Overfitting is addressed by reducing the number of parameters via depth-wise separable convolution. Compared to 2D convolution, depth-wise convolution reduces the encoder's parameters by 4.3% and the decoder's by 3%, with a total parameter reduction of 15.9%, without any painful trade-off for feature learning, efficiently reducing the parameters in the spatial feature extraction layers (i.e., Layers 1 and 2) by a factor of n. This makes the model sufficiently lightweight and efficient for real time anomaly detection.

Secondly, the use of the original SVD_{loss} in the generator allows us to minimise the distance between low-rank optimal approximations of input and generated images, aiding network convergence. As a result our SVD-GAN model starts converging as early as the 167th epoch (with LeakyReLU as normaliser), gradually attaining a stable reconstruction characterised by low error values for 'normal' frames.

Figure 4 illustrates the anomaly detection graph, with highlighted the frames which are false positives and true positives, for a sequence of the ShanghaiTech dataset. The robustness of the proposed system's reconstruction abilities is visually illustrated in Fig. 5. More examples of reconstructed video frames are shown in Fig. S2 of the Supplementary material.

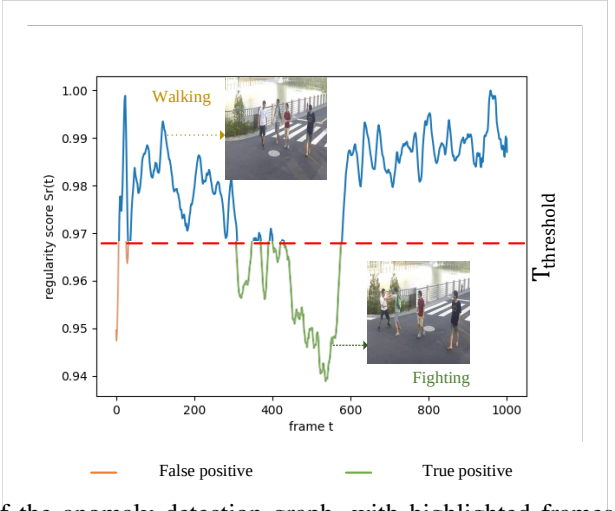

Figure 4: Plot of the anomaly detection graph, with highlighted frames which are false positives and true positives, for a test sequence of the ShanghaiTech dataset.

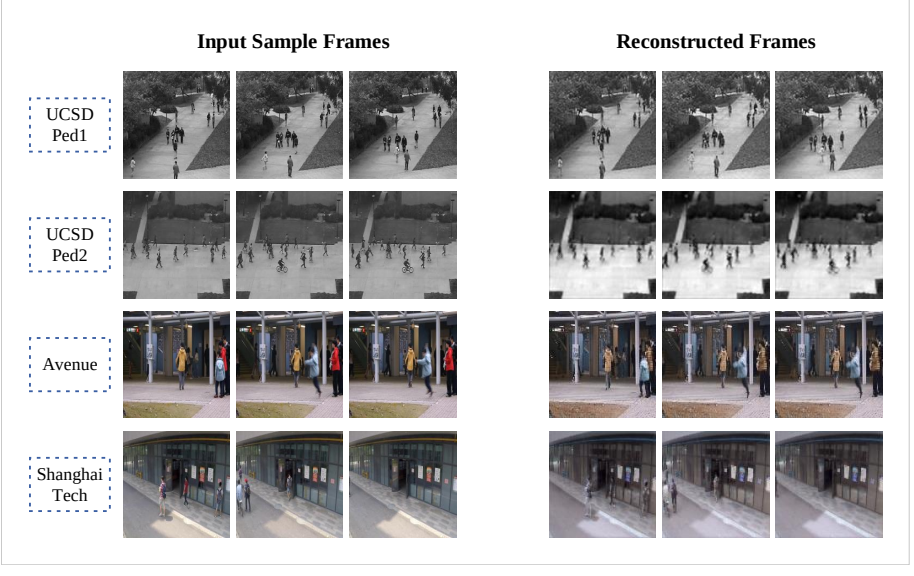

Figure 5: Examples of frames reconstructed by SVD-GAN, compared to the original inputs.

5 Conclusions

The proposed lightweight SVD-GAN architecture has a clear edge over state-of-art unsupervised anomaly detection methods while using fewer parameters, thanks to the use of temporal blocks for better spatiotemporal feature learning and an original SVD loss for more robust GAN learning. Our experiments show that our system widely outperforms prior art on both the Avenue and ShanghaiTech datasets and can leverage large-scale datasets. In the future, the accuracy of the system can be further improved by using a memory module or a state-of-the-art 3D feature extractor.

References

[1] Integrating prediction and reconstruction for anomaly detection. *Pattern Recognition Letters*, 129:123–130, 2020. ISSN 0167-8655. doi: https://doi.org/10.1016/j.patrec.2019.11.024.

[2] Samet Akcay, Amir Atapour-Abarghouei, and Toby P. Breckon. Ganomaly: Semi-supervised anomaly detection via adversarial training. In C. V. Jawahar, Hongdong Li, Greg Mori, and Konrad Schindler, editors, *Computer Vision – ACCV 2018*, pages 622–637, Cham, 2019. Springer International Publishing.

[3] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan, 2017.

[4] D. V. S. Chandra. Digital image watermarking using singular value decomposition. In *The 2002 45th Midwest Symposium on Circuits and Systems, 2002. MWSCAS-2002.*, volume 3, pages III–III, 2002. doi: 10.1109/MWSCAS.2002.1187023.

[5] F. Chen and M. R. Jahanshahi. Nb-cnn: Deep learning-based crack detection using convolutional neural network and naïve bayes data fusion. *IEEE Transactions on Industrial Electronics*, 65(5):4392–4400, 2018. doi: 10.1109/TIE.2017.2764844.

[6] MyeongAh Cho, Taeoh Kim, and Sangyoun Lee. Unsupervised video anomaly detection via flow-based generative modeling on appearance and motion latent features. *CoRR*, abs/2010.07524, 2020. URL https://arxiv.org/abs/2010.07524.

[7] Yong Shean Chong and Yong Haur Tay. Abnormal event detection in videos using spatiotemporal autoencoder. In Fengyu Cong, Andrew Leung, and Qinglai Wei, editors, *Advances in Neural Networks - ISNN 2017*, pages 189–196, Cham, 2017. Springer International Publishing.

[8] Jeff Donahue, Philipp Krähenbühl, and Trevor Darrell. Adversarial feature learning. *ICLR*, 2017.

[9] Maayan Frid-Adar, Idit Diamant, Eyal Klang, Michal Amitai, Jacob Goldberger, and Hayit Greenspan. Gan-based synthetic medical image augmentation for increased cnn performance in liver lesion classification. *Neurocomputing*, 321:321–331, 2018. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2018.09.013. URL https://www.sciencedirect.com/science/article/pii/S0925231218310749.

[10] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014.

[11] M. Hasan, J. Choi, J. Neumann, A. K. Roy-Chowdhury, and L. S. Davis. Learning temporal regularity in video sequences. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 733–742. IEEE Computer Society, jun 2016. doi: 10.1109/CVPR.2016.86.

[12] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9:1735–80, 12 1997. doi: 10.1162/neco.1997.9.8.1735.

[13] P. Isola, J. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5967–5976. IEEE Computer Society, jul 2017. doi: 10.1109/CVPR.2017.632. URL https://doi.ieeecomputersociety.org/10.1109/CVPR.2017.632.

[14] B Ravi Kiran, Dilip Mathew Thomas, and Ranjith Parakkal. An overview of deep learning based methods for unsupervised and semi-supervised anomaly detection in videos. *Journal of Imaging*, 4(2):36, 2018.

[15] Jin-Hwan Lee, Sung-Sik Yoon, In-Ho Kim, and Hyung-Jo Jung. Diagnosis of crack damage on structures based on image processing techniques and R-CNN using unmanned aerial vehicle (UAV). In Hoon Sohn, editor, *Sensors and Smart Structures Technologies for Civil, Mechanical, and Aerospace Systems 2018*, volume 10598, pages 265 – 272. International Society for Optics and Photonics, SPIE, 2018. doi: 10.1117/12.2296691. URL https://doi.org/10.1117/12.2296691.

[16] Sangmin Lee, Hak Gu Kim, and Yong Man Ro. Stan: Spatio-temporal adversarial networks for abnormal event detection. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1323–1327, 2018. doi: 10.1109/ICASSP.2018.8462388.

[17] W. Li, V. Mahadevan, and N. Vasconcelos. Anomaly detection and localization in crowded scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36 (1):18–32, 2014. doi: 10.1109/TPAMI.2013.111.

[18] Zachary C Lipton and Subarna Tripathi. Precise recovery of latent vectors from generative adversarial networks. *arXiv preprint arXiv:1702.04782*, 2017.

[19] Wen Liu, Weixin Luo, Dongze Lian, and Shenghua Gao. Future frame prediction for anomaly detection - a new baseline. pages 6536–6545, 06 2018. doi: 10.1109/CVPR.2018.00684.

[20] C. Lu, J. Shi, and J. Jia. Abnormal event detection at 150 fps in matlab. In *2013 IEEE International Conference on Computer Vision*, pages 2720–2727, 2013. doi: 10.1109/ICCV.2013.338.

[21] W. Luo, W. Liu, and S. Gao. A revisit of sparse coding based anomaly detection in stacked rnn framework. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 341–349, 2017. doi: 10.1109/ICCV.2017.45.

[22] W. Luo, W. Liu, and S. Gao. Remembering history with convolutional lstm for anomaly detection. In *2017 IEEE International Conference on Multimedia and Expo (ICME)*, pages 439–444, Los Alamitos, CA, USA, jul 2017. IEEE Computer Society. doi: 10.1109/ICME.2017.8019325. URL https://doi.ieeecomputersociety.org/10.1109/ICME.2017.8019325.

[23] Vijay Mahadevan, Weixin Li, Viral Bhalodia, and Nuno Vasconcelos. Anomaly detection in crowded scenes. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1975–1981. IEEE, 2010.

[24] J. Medel. Anomaly detection using predictive convolutional long short-term memory units. 2016.

[25] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *CoRR*, abs/1411.1784, 2014. URL http://arxiv.org/abs/1411.1784.

[26] Marc Moonen, Paul Van Dooren, and Joos Vandewalle. A singular value decomposition updating algorithm for subspace tracking. *SIAM Journal on Matrix Analysis and Applications*, 13(4):1015–1038, 1992.

[27] Trong Nguyen Nguyen and Jean Meunier. Anomaly detection in video sequence with appearance-motion correspondence. In *2019 IEEE/CVF (ICCV)*, pages 1273–1283, 2019. doi: 10.1109/ICCV.2019.00136.

[28] Viorica Patraucean, Ankur Handa, and Roberto Cipolla. Spatio-temporal video autoencoder with differentiable memory. *CoRR*, abs/1511.06309, 2015. doi: https://doi.org/10.17863/CAM.26485.

[29] A. Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *CoRR*, abs/1511.06434, 2016.

[30] Mahdyar Ravanbakhsh, Moin Nabi, Enver Sangineto, Lucio Marcenaro, Carlo Regazzoni, and Nicu Sebe. Abnormal event detection in videos using generative adversarial nets. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 1577–1581, 2017. doi: 10.1109/ICIP.2017.8296547.

[31] Rowayda A Sadek. Svd based image processing applications: state of the art, contributions and research challenges. *arXiv preprint arXiv:1211.7102*, 2012.

[32] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 2234–2242, Red Hook, NY, USA, 2016. Curran Associates Inc. ISBN 9781510838819.

[33] Dinesh Jackson Samuel and Fabio Cuzzolin. Unsupervised anomaly detection for a smart autonomous robotic assistant surgeon (saras) using a deep residual autoencoder. *IEEE Robotics and Automation Letters*, 6(4):7256–7261, 2021. doi: 10.1109/LRA.2021.3097244.

[34] T. Schlegl, Philipp Seeböck, S. Waldstein, G. Langs, and U. Schmidt-Erfurth. f-anogan: Fast unsupervised anomaly detection with generative adversarial networks. *Medical Image Analysis*, 54:30–44, 2019.

[35] Thomas Schlegl, Philipp Seeböck, Sebastian M. Waldstein, Ipek Oguz, Pew-Thian Yap, and Dinggang Shen. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In *Information Processing in Medical Imaging*, pages 146–157, Cham, 2017. Springer International Publishing.

[36] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wangchun Woo. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In *Advances in Neural Information Processing Systems 28: Annual*

Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada, pages 802–810, 2015.

[37] Dinesh Singh and C. Krishna Mohan. Graph formulation of video activities for abnormal activity recognition. *Pattern Recogn.*, 65(C):265–272, May 2017. ISSN 0031-3203. doi: 10.1016/j.patcog.2017.01.001. URL https://doi.org/10.1016/j.patcog.2017.01.001.

[38] Angela A. Sodemann, Matthew P. Ross, and Brett J. Borghetti. A review of anomaly detection in automated surveillance. *IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews*, 42(6):1257–1272, December 2012. ISSN 1094-6977. doi: 10.1109/TSMCC.2012.2215319.

[39] H. Song, C. Sun, X. Wu, M. Chen, and Y. Jia. Learning normal patterns via adversarial attention-based autoencoder for abnormal event detection in videos. *IEEE Transactions on Multimedia*, 22(8):2138–2148, 2020. doi: 10.1109/TMM.2019.2950530.

[40] W. Sultani, C. Chen, and M. Shah. Real-world anomaly detection in surveillance videos. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6479–6488, 2018. doi: 10.1109/CVPR.2018.00678.

[41] Hung Vu, Tu Dinh Nguyen, Trung Le, Wei Luo, and Dinh Phung. Robust anomaly detection in videos using multilevel representations. In Pascal Van Hentenryck and Zhi-Hua Zhou, editors, *Proceedings of AAAI19-Thirty-Third AAAI conference on Artificial Intelligence*, number 1 in Proceedings of the AAAI Conference on Artificial Intelligence, pages 5216–5223, United States of America, 2019. Association for the Advancement of Artificial Intelligence (AAAI). doi: 10.1609/aaai.v33i01.33015216. URL https://aaai.org/Conferences/AAAI-19/. AAAI Conference on Artificial Intelligence 2019, AAAI 2019 ; Conference date: 27-01-2019 Through 01-02-2019.

[42] Xuanzhao Wang, Zhengping Che, Bo Jiang, Ning Xiao, Ke Yang, Jian Tang, Jieping Ye, Jingyu Wang, and Qi Qi. Robust unsupervised video anomaly detection by multipath frame prediction. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–12, 2021. doi: 10.1109/TNNLS.2021.3083152.

[43] D. Weimer, B. Scholz-Reiter, and M. Shpitalni. Design of deep convolutional neural network architectures for automated feature extraction in industrial inspection. *Cirp Annals-manufacturing Technology*, 65:417–420, 2016.

[44] Houssam Zenati, Chuan Sheng Foo, Bruno Lecouat, Gaurav Manek, and Vijay Ramaseshan Chandrasekhar. Efficient gan-based anomaly detection, 2018. URL https://openreview.net/forum?id=BkXADmJDM.

[45] Y. Zhang, H. Lu, L. Zhang, and X. Ruan. Combining motion and appearance cues for anomaly detection. *Pattern Recognit.*, 51:443–452, 2016.

[46] B. Zhao, Li Fei-Fei, and E. Xing. Online detection of unusual events in videos via dynamic sparse coding. *CVPR 2011*, pages 3313–3320, 2011.

[47] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34 (07):13001–13008, Apr. 2020. doi: 10.1609/aaai.v34i07.7000. URL https://ojs.aaai.org/index.php/AAAI/article/view/7000.